# State of the Art Reports

## Natural Language Translation

### Computer-based translation systems and tools

John Hutchins

Machine translation (MT) is still better known for its failures than for its successes. It continues to labour under misconceptions and prejudices from the ALPAC report of more than thirty years ago, and now it has to contend with widespread misunderstanding and ridicule from users of online MT services. The goal of developing fully automatic general-purpose systems capable of near-human translation quality has been long abandoned. The aim is now to produce aids and tools for professional and non-professional translation which exploit the potentials of computers to support human skills and intelligence, or which provide rough translations for users to extract the essential information from texts in foreign languages. The following brief survey covers commercial and research developments in the field, with an emphasis on European developments.

### The recent growth of MT

The traditional MT user has been the large multinational company, requiring technical documentation and operating manuals in a range of languages. The system runs on a mainframe and produces 'raw' output of variable quality for revision (post-editing) by translators. A successful alternative has been the pre-editing of input texts (typically with a controlled 'regularised' language) to minimise the expensive editing processes. Both these types of MT use are continuing to expand rapidly. There are now millions of pages of translation produced in this way every year.

Although MT software for personal computers began to appear in the early 1980s (with the Weidner MicroCAT system becoming particularly successful), it was during the 1990s that sales of these systems showed a dramatic rise. There are now estimated to be more than 1000 different MT packages on sale (when each language pair is counted separately.) Some products have had large sales, e.g. Globalink and those from Lernout & Hauspie, and in Japan one system (Korya Eiwa from Catena, for English-Japanese translation) was said to have sold over 100,000 copies in its first year on the market. Nearly all the Japanese computer and software companies seem to have a product (Fujitsu, Toshiba, Brother, Mitsubishi, Sharp, Hitachi, NEC, Kodensha, Nova, Oki, etc.), mainly for Japanese and English in both directions. Outside Japan, products come mostly from independent companies set up to develop and market a range of translation software products (e.g. AppTek, CITAC, LogoMedia, Linguatec, etc.). For details of systems see the “Compendium of translation software” on the EAMT website (http://www.eamt.org/compendium.html)

Though it is difficult to establish how much of the cheaper software purchased is regularly used (some cynics claim that only a very small proportion is tried out more than once), there is no doubting the growing volume of 'occasional' translation, i.e. by people from all backgrounds wanting gists of foreign text in their own language, or wanting to communicate in writing with others in other languages, however poor the quality. This latent market for low-quality translation was untapped until relatively recently, but is now being exploited on a large scale.

As well as ‘home’ users, other users of PC products are the independent professional translators. Many vendors of older mainframe systems (Systran, Fujitsu, Metal, Logos) have downsized their systems, retaining most features of their mainframe products. Systran Personal and Systran Premium, for example, are Windows-based versions of the successful system developed since the 1960s for clients worldwide in a large range of languages. Systran Personal is aimed at the ‘home’ user and Systran Premium for the professional independent translator. Both are available for the language pairs (in both directions) English-French, English-German, English-Italian, English-Portuguese,

English-Spanish and for French-German and French-Spanish. Linguatec markets an English-German system (Personal Translator PT) originally developed by IBM for a mainframe system, and the publishing company Langenscheidt markets a PC version of the old Metal system (T1, in conjunction with Sail Labs) for German-English, German-French, German-Spanish, and German-Russian (in both directions) in versions for 'home' use and for professionals.

At the same time, many MT vendors are providing network-based translation services for on-demand translation, with human revision as optional extras. CompuServe led the way some eight years ago, with on-line translation of electronic mail. Shortly afterwards, Systran made its systems available online in the Babelfish service of AltaVista – both for text translation and for web page translation. Many other services have followed (FreeTranslation, Gist-in-Time, Reverso, ProMT, etc.) In addition, of course, the demand for translation of Internet information has resulted in most stand-alone PC software products also being now capable of translating downloaded Web pages and electronic mail messages.

With cheaper PC software and wider access to the Internet, there has undoubtedly been an unprecedented growth in the use of fully automatic systems, primarily among non-professional translators for the production of rough versions. However, these are often not appropriate for professional translators, translation agencies and many companies. Here the preference may well be the use of translator workstations.

## Translator workstations

For professional translators, the attraction of the workstation is the integration of tools from simple word processing aids (spelling and grammar checkers) to full automatic translation. The translator can choose to make use of whichever tool seems most appropriate for the task in hand. The vendors of these systems always stress that translators do not have to change their work patterns; the systems aim to increase productivity with translator-oriented tools which are easy to use and fully compatible with existing word processing systems.

The three most widely used workstations all originate from Europe: Trados' Translation Workbench, STAR's Transit, and Atril's Déjà Vu. In facilities and functions, each offer similar ranges: multilingual split-screen word processing, terminology recognition, retrieval and management, translation memory (pre-translation based on existing texts), alignment software for users to create their own bilingual text databases, retention of original text formatting, and support for a very wide range of European languages, both as source and target languages. For translators, one of the most useful facility is storage and access to previously translated texts (their own or the company's) in a 'translation memory', enabling them to avoid the re-translation of unchanged texts or to extract and adapt examples of previous translations. In addition, most workstations include integrated access to MT systems, for translators to use as further assistance.

## MT in Europe

Most of the PC-based MT software originates from Japan and the United States, and sales have been lower in Europe. However, there are notable European products: the Comprendium and T1 systems (Sail Labs), Personal Translator PT (Linguatec), the iTranslator series (originally Lernout & Hauspie, now Bowne Global), the Reverso systems (Softissimo), the range of ProMT systems (for Russian to/from English and German); and the PARS systems for Russian and Ukrainian to/from English. Other PC-based systems from Europe include: Hypertrans for translating between English, French, German, Italian, Spanish and Portuguese; PeTra for translating between Italian and English; the Al-Nakil system for Arabic, French and English; the Winger system for English to/from Danish, French and Spanish; and the TranSmart system for Finnish-English from Kielikone Ltd. Most of these systems are available in different versions for large enterprises, for independent professional translators, and for occasional (home) use, e.g. for translating Web pages and emails.

As the methods and techniques of natural language processing become more familiar outside research laboratories, many companies have been developing their own translation tools. Although a worldwide trend (e.g. PAHO and Smart in the US, and NHK, JICST and CSK in Japan), it is a distinctive feature of European MT activity. Custom-built MT is the speciality of Cap Volmac Lingware Services, a Dutch subsidiary of the Cap Gemini Sogeti Group, which has constructed controlled-language systems for textile and insurance companies, mainly from Dutch to English. Another company developing custom-built MT systems with controlled languages is Xplanation b.v. (previously LANT) in Louvain, Belgium, using software developed for the METAL system. One of the best known success story for custom-built MT is the PaTrans system developed for LingTech A/S to translate English patents into Danish. The system is based on methods and experience gained from the Eurotra project of the European Commission.

In Europe, the main users of MT systems and translation tools have been large translation services and multinational companies, e.g. the software group SAP is translating some 8 millions words a year; and the European Commission has seen a rapid growth in the use of Systran, now some million pages a year. The Commission's Translation Service has developed its own workstation, EURAMIS, optimizing use of its linguistic resources, the Eurodicautom and CELEX databases, and collections of translated European Union documents (as a 'translation memory'), and providing easy access to its own MT system and other MT services.

The most distinctive feature of the European scene is the growth of companies providing software localisation, and which are thus acquiring considerable experience in the use of translation aids and MT systems. Computer-based support tools for localisation processes, including quality assurance and project management, are now widely available as well as systems designed specifically for web localisation (e.g. IBM's WebSphere). A forum for the interchange of experience and the establishment of standards was set up in 1990: the Localisation Industry Standards Association, publishing a newsletter ('LISA Forum'). Ireland, as the main centre for these services, has its own Software Localisation Group and has set up a Localisation Resources Centre (with support from the Irish government and EU.)

## MT research

As already indicated, nearly all the major computer software companies are showing interest in developing translation tools and systems, with Japanese and American companies in the vanguard; the growth in sales of PC-based products has opened up a huge potential market.

By comparison, academic research has declined in the area of written language MT as such (i.e. as distinct from speech translation and multilingual applications in language engineering). Nevertheless, there is research on developing dialogue-based systems which combine computer-interactive authoring and translating into an unknown language, usually within a restricted subject field in order to ensure higher quality output. There is much interest in exploring new techniques in neural networks, parallel processing, and particularly in corpus-based approaches to MT: statistics-based systems using little or no linguistic data, example-based MT systems generating texts from fragments extracted from already translated texts, hybrid systems combining traditional linguistic rules and statistical methods, and so forth. Above all, the crucial problem of lexicon acquisition (always a bottleneck for MT) is receiving major attention by many academic research groups, and the large lexical and text resources (such as those available from the Linguistic Data Consortium and ELRA) are being widely and fruitfully exploited. University MT research groups are increasingly working jointly with commercial organisations to develop customer-specific systems, e.g. Carnegie-Mellon University and the Caterpillar Corporation, or to undertake basic research for companies.

However, the main emphasis of research has shifted to applications within the context of multilingual tools for specific needs and to longer-term research on speech translation. Since the ending of Eurotra, research funds from the European Union have been more widely focused on projects within the broad field of language engineering, which includes

multilingual tools of all kinds as well as translation assistance in various contexts. Practical implementation and collaboration with industrial partners is emphasised throughout, as well as the need for general-purpose and re-usable products. Very many of these multilingual projects involve translation of some kind, usually within a restricted subject field and often in controlled conditions.

The planned accession of states in Central and Eastern Europe to the European Union has stimulated research on MT and translation tools for languages such as Czech, Polish, Hungarian, Slovenian, and Estonian – not just for supporting translation of treaty and other legal documents but also for enabling public access to information resources. Indeed, today many projects funded by the European Union within the broad field of human language technology (www.cordis.lu/ist/) involve multilingual tools of all kinds and include translation aids, usually within a restricted subject field and often in controlled conditions, and many are designed for Internet applications. Mention should also be made of research on systems for 'minority' languages in Europe, such as Basque, Catalan and Galician in Spain and immigrant languages such as Hindi, Bengali and Gujarati in the United Kingdom. The need is both for full translation systems and for translation aids, dictionaries, glossaries, bilingual corpora of authorized translations, etc.

## Speech translation

The goal of automatic speech translation was always a distant vision for MT, until developments in speech technology began to make it a feasible objective from the 1980s. At first, research was on a small scale, e.g. the project at British Telecom to translate formulaic messages over the telephone. Later, in Japan a joint government and industry company ATR was established in 1986 near Osaka, and is now one of the main centres for automatic speech translation. The aim is to develop a speaker-independent real-time telephone translation system for Japanese to English and vice versa, initially for hotel reservation and conference registration transactions.

Other speech translation projects were set up subsequently. The JANUS system is a research project at Carnegie-Mellon University and at Karlsruhe in Germany. The researchers are collaborating with ATR in a consortium (C-STAR), each developing speech recognition and synthesis modules for their own languages (English, German, Japanese).

In May 1993 the Verbmobil project was funded by the German Ministry for Research and Technology. The aim was a portable translation aid for business negotiations to supplement users' own knowledge of the languages (German, Japanese, English). Numerous German university groups were involved in fundamental research on dialogue linguistics, speech recognition and MT design; and although the main goal was not achieved at the close of the project in 2000 the development of efficient methodologies for dialogue and speech translation and the establishment of top-class research groups in Germany are regarded as notable successes.

Speech translation is probably at present the most innovative area of computer-based translation research, and it is attracting most funding and the most publicity. However, few experienced observers expect dramatic developments in this area in the near future -- the development of MT has taken many years to reach the present stage: widespread practical use in multinational companies, a wide range of PC-based products of variable quality and application, growing use on networks and for electronic mail and web pages; and researchers know that there is still much to be done to improve quality.



For later information please see the author's personal website at http://ourworld.compuserve.com/homepages/wjhutchins and also the EAMT website at http://www.eamt.org, where details of current commercial MT systems and translation aids can be found in its "Compendium of translation software".